Students Performance Prediction

Mentor: Prof. Rakesh Jaiswal

Author: Aarya Sharma, Akanksha Verma, Akriti Gupta,Priyanshi Gupta

Institution :Oriental Institute of Science and Technology

1.Abstract

Student performance prediction is essential for early identification of struggling students and personalized learning interventions. This research presents a Machine Learning (ML)-based approach to predict student performance using factors like attendance, previous exam grades, study hours, motivation, and assignment completion. We implemented Decision Tree, Random Forest, and Logistic Regression models to classify students into performance levels (Low, Medium, High) or Pass/Fail categories. The dataset was collected through Google Forms and pre-processed to handle missing values and normalize numerical data. The results indicate that Random Forest provides the best accuracy, and attendance and previous exam scores are the most influential factors. The findings demonstrate how ML can help educators track student progress and offer personalized academic support.

2.Keywords:

Student Performance Prediction, Machine Learning in Education, Decision Tree Classification, Random Forest Algorithm, Academic Success Factors, Predictive Analytics in Education Student Success Prediction, Attendance and Academic Performance Supervised Learning Models, Study Hours and Grades Correlation AI in Education, Early Intervention for Students, Personalized Learning Using ML.

3.Introduction:

Student success depends on multiple academic and behavioural factors such as study habits, motivation, attendance, and engagement with assignments. Traditional student performance tracking methods, like teacher assessments and periodic tests, fail to identify struggling students early. This often results in delayed interventions, affecting overall academic success.

With the rise of data-driven technologies, Machine Learning (ML) offers a systematic approach to predicting student performance by analysing past trends. By leveraging ML, we can detect patterns in student behaviour and provide data-driven insights, allowing educators to take proactive measures.

3.1 Problem Statement

Traditional methods of performance evaluation:

- Rely on limited data (mainly exam scores).
- Fail to provide early intervention for struggling students.
- Are subjective and differ from teacher to teacher.

A Machine Learning-based prediction system can analyse multiple factors and identify at-risk students before their performance deteriorates.

3.2 Objectives

The key objectives of this research are:

1. To develop an ML model that predicts student performance based on academic and behavioural factors.
2. To identify key parameters influencing student success.

3. To analyse ML model performance and determine the best algorithm for student prediction.
4. To provide educators with insights that help personalize teaching strategies.

4. Literature Review

Several studies have explored Machine Learning in education, focusing on student performance prediction. Some key findings include:

- Decision Trees and Random Forest models have been widely used due to their interpretability and efficiency in handling categorical and numerical data.
- Neural Networks offer higher accuracy but require larger datasets.
- Attendance and previous exam scores have been identified as strong predictors of academic performance in multiple studies.
- Research has also highlighted the need for more behavioural data (motivation, stress levels, extracurricular activities) to improve model accuracy.

This study builds on these insights by implementing Decision Tree, Random Forest, and Logistic Regression models on a small-scale dataset collected through surveys

5.Methodology

5.1 Dataset Collection

Data was collected through Google Forms, where students entered responses for the following parameters:

- Attendance (%)
- Previous Exam Grades
- Study Hours per Day
- Motivation Level (Scale: 1-5)
- Assignment Completion Rate (%)

The collected data was stored in CSV format for processing.

5.2 Data Preprocessing

Before training the ML models, the dataset was cleaned and pre-processed:

- Handling Missing Values:

If attendance or grades were missing, they were replaced with the mean value.

- Feature Scaling:

Study hours and grades were normalized to maintain uniformity.

- Encoding Categorical Data:

Motivation levels (Low, Medium, High) were converted into numerical values(1-5).

- Splitting the Dataset:

The data was divided into 80% training and 20% testing sets.

5.3 Machine Learning Models Implemented

Three models were implemented:

Algorithm	Why Used?	Application in the Project
Decision Tree	Simple, interpretable, works well for small datasets	Classifying Students into Low, Medium, High performance
Random Forest	More accurate than a single Decision Tree, reduces overfitting	Predicts student outcomes more reliably
Logistic Regression	Best for binary classification tasks	Used to predict Pass/Fail status

5.4 Model Training & Testing

The models were trained using Scikit-Learn.

Performance was evaluated using:

- Accuracy (overall correctness of predictions).
- Precision, Recall, and F1-score (measuring model effectiveness).

6. Results & Discussion

Once the models were trained and tested, we evaluated their performance based on accuracy and key metrics like Precision, Recall, and F1-score. Below are the detailed results and insights obtained from the study.

6.1 Model Performance Comparison

The following table compares the accuracy of three different machine learning models used in the project:

Algorithm	Accuracy (%)	Precision	Recall	F1-score
Decision Tree	82%	0.81	0.79	0.80
Random Forest	87%	0.85	0.84	0.85
Logistic Regression	75%	0.72	0.70	0.71

Key Findings:

- ♦ Random Forest performed the best, achieving the highest accuracy of 87%.
- ♦ Decision Tree was slightly less accurate (82%), but it remains the easiest to interpret.
- ♦ Logistic Regression had the lowest accuracy (75%), making it less suitable for multi-category classification (Low, Medium, High).
- ♦ Precision, Recall, and F1-score were highest for Random Forest, confirming its effectiveness in handling small student datasets.

6.2 Key Insights & Feature Importance

To determine which factors had the greatest influence on student performance, we analyzed feature importance scores from the Random Forest model. The most significant parameters were:

Feature	Importance (%)
Previous Exam Grades	38%
Attendance	30%
Study Hours Per Day	18%
Motivation Level	10%
Assignment Completion Rate	4%

Key Takeaways:

- ♦ Previous exam grades were the strongest predictor of future performance (38% impact), meaning students with consistently good academic records are more likely to continue performing well.
- ♦ Attendance played a crucial role (30%), reinforcing the importance of class participation in academic success.
- ♦ Study hours (18%) had a direct impact, but some students still performed well despite fewer study hours (suggesting differences in study quality).
- ♦ Motivation levels (10%) showed moderate correlation, indicating that personal drive and engagement influence academic outcomes.
- ♦ Assignment completion had the least impact (4%), possibly because different teachers assign different weights to assignments.

7.Impact & Benefits

The implementation of a Machine Learning-based student performance prediction model offers several benefits to students, educators, and institutions.

7.1 Impact of the Project

1. For Students:
- Helps them understand their strengths and weaknesses.
- Provides personalized feedback on study habits and attendance.
- Encourages self-improvement through data-driven recommendations.

2. For Educators:
- Identifies struggling students early and enables intervention.
- Helps teachers design customized learning plans.
- Allows tracking of attendance and engagement trends.

3. For Institutions:
- Supports data-driven decision-making in academic planning.
- Improves student retention rates by addressing performance gaps.
- Helps create an early warning system to prevent academic failures.

7.2 Benefits of the Project

- ♦ Early Identification of At-Risk Students: The ML model can flag students who need additional academic support, allowing teachers to intervene before final exams.

♦ Personalized Learning & Recommendations: By analyzing individual study habits and motivation levels, educators can suggest tailored study strategies.
♦ Automated & Data-Driven Evaluation: Unlike traditional methods, this approach reduces subjectivity in student performance analysis.

♦ Scalability for Larger Student Groups: Although this project focuses on a small-scale dataset, it can be expanded for larger institutions with additional data sources (e.g., behavioral analysis, extracurricular activities).

♦ Potential for Real-Time Dashboard Integration: If deployed as a web application, schools/universities can monitor student progress dynamically.

8. Conclusion & Future Scope

8.1 Conclusion

1. Machine Learning successfully predicts student performance based on attendance, exam scores, study habits, and motivation.
2. Random Forest was the best-performing model, demonstrating its ability to capture multiple factors accurately.
3. ML models can help institutions and educators implement early intervention strategies.

8.2 Future Scope

♦ Expanding the dataset to include psychological and behavioral factors like stress levels, peer influence, and extracurricular activities.
♦ Implementing deep learning models (Neural Networks) for higher accuracy.
♦ Integrating the model into a web application for real-time student performance tracking.

8.3. References

1. **Orji, F. A., & Vassileva, J. (2022).**
Machine Learning Approach for Predicting Students' Academic Performance and Study Strategies Based on Their Motivation.
2. **Agyemang, E. F., Mensah, J. A., Ampomah, O.-A., Agyekum, L., Akuoko-Frimpong, J., Quansah, A., & Akinlosotu, O. M. (2024).**
Predicting Students' Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application.
3. **Chen, Y., Sun, J., Wang, J., Zhao, L., Song, X., & Zhai, L. (2025).**
Machine Learning-Driven Student Performance Prediction for Enhancing Tiered Instruction.
4. **Baker, R. S. J. d., & Inventado, P. S. (2014).**
Educational Data Mining and Learning Analytics.
5. **Romero, C., & Ventura, S. (2010).**
Data Mining in Education.
6. **Peña-Ayala, A. (Ed.). (2014).**
Educational Data Mining: Applications and Trends.
7. **John MacGregor**
Predictive Analytics in Education: Applications and Trends.